Exploratory Data Analysis

Lecture 4 & 5

Corina Besliu

Technical University of Moldova

September 16, 2021

1. Graphical Techniques for Interval Data Continued...
   - The Ogive Curve
   - Box Plots
   - Line Charts

2. Description of the Relationship Between Two Variables
   - Scatter Plot
   - Covariance
   - Correlation Coefficient

3. Practical Assignments in Python

Telephone Bills Frequency Distribution

To build the histogram and the Stem-and-Leaf Charts we used the Frequency Distribution

CLASS LIMITS	FREQUENCY
0 to 15*	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
105 to 120	14
Total	200

Relative Frequency Distribution

To learn the proportion of the observations that fall into each class create the Relative Frequency Distribution.

CLASS LIMITS	RELATIVE FREQUENCY
0 to 15	71/200 = .355
15 to 30	37/200 = .185
30 to 45	13/200 = .065
45 to 60	9/200 = .045
60 to 75	10/200 = .050
75 to 90	18/200 = .090
90 to 105	28/200 = .140
105 to 120	14/200 = .070
Total	200/200 = 1.0

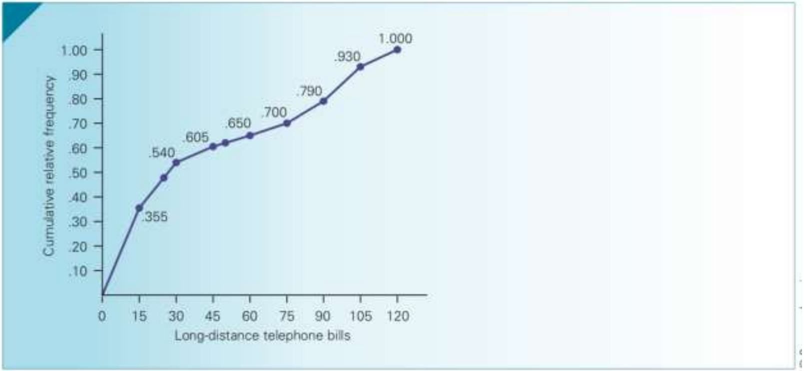
Cumulative Relative Frequency Distribution

Often with numerical(interval) data we may be interested in knowing the percentage of observations that lie bellow a certain value. To Learn this use the cumulative frequency distribution.

CLASS LIMITS	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
0 to 15	71/200 = .355	71/200 = .355
15 to 30	37/200 = .185	108/200 = .540
30 to 45	13/200 = .065	121/200 = .605
45 to 60	9/200 = .045	130/200 = .650
60 to 75	10/200 = .05	140/200 = .700
75 to 90	18/200 = .09	158/200 = .790
90 to 105	28/200 = .14	186/200 = .930
105 to 120	14/200 = .07	200/200 = 1.00

The Ogive Plot

To visualize the cumulative frequency distribution use the Ogive Curve.



Percentiles and Measures of Relative Standing

Percentiles are mesures of relative standing. They give you an idea about the position of particular values relative to the entire data set (e.g. median).

Percentile

The *P*th **percentile** is the value for which *P* percent are less than that value and $(100 - P)\%$ are greater than that value.

Example: The Graduate Management Admission Test (GMAT) is always reported with information of your relative standing, "your score xxx is in the 50th percentile"

Percentiles and Measures of Relative Standing

Percentiles are mesures of relative standing. They give you an idea about the position of particular values relative to the entire data set (e.g. median).

Percentile

The Pth **percentile** is the value for which P percent are less than that value and $(100 - P)\%$ are greater than that value.

Example: The Graduate Management Admission Test (GMAT) is always reported with information of your relative standing, "your score xxx is in the 50th percentile"

Percentiles and Quartiles

The 25th, 50th, and 75th percentiles are also called Quartiles. Because they divide the set of data into quarters.

- The 1st or lower quartile: Q1 = 25th percentile
- The 2nd quartile: Q2 = 50th percentile = median
- The 3rd or upper quartile: Q3 = 75th percentile

Percentiles and Quartiles

The 25th, 50th, and 75th percentiles are also called Quartiles. Because they divide the set of data into quarters.

- The 1st or lower quartile: Q1 = 25th percentile
- The 2nd quartile: Q2 = 50th percentile = median
- The 3rd or upper quartile: Q3 = 75th percentile

Percentiles and Quartiles

The 25th, 50th, and 75th percentiles are also called Quartiles. Because they divide the set of data into quarters.

- The 1st or lower quartile: $Q1 = 25\text{th percentile}$
- The 2nd quartile: $Q2 = 50\text{th percentile} = \text{median}$
- The 3rd or upper quartile: $Q3 = 75\text{th percentile}$

Percentiles and Quartiles

The 25th, 50th, and 75th percentiles are also called Quartiles. Because they divide the set of data into quarters.

- The 1st or lower quartile: $Q1 = 25\text{th percentile}$
- The 2nd quartile: $Q2 = 50\text{th percentile} = \text{median}$
- The 3rd or upper quartile: $Q3 = 75\text{th percentile}$

The Interquartile Range

The interquartile range measures the spread of the middle 50% of the observations.

Interquartile Range

$$\text{Interquartile range} = Q_3 - Q_1$$

Question: What does a high value of the Interquantile Range tell us about the Variance and Standard Deviation of this distribution?

The Interquartile Range

The interquartile range measures the spread of the middle 50% of the observations.

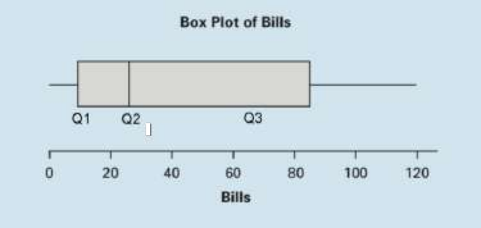
Interquartile Range

$$\text{Interquartile range} = Q_3 - Q_1$$

Question: What does a high value of the Interquantile Range tell us about the Variance and Standard Deviation of this distribution?

The Box Plot

The Box Plot shows us 5 statistics.



Q1 = 9.275 Q2 = 26.905 Q3 = 84.9425

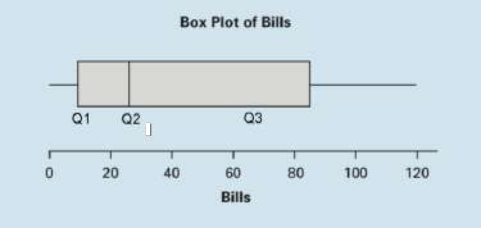
Interquartile range: Q3 - Q1 = 75.6675

To find the outliers:

- 1.5 x 75.6675 = 113.5013
- 9.275 - 113.5013 = 104.226
- 84.9425 + 113.5013 = 198.4438.

The Box Plot

The Box Plot shows us 5 statistics.



Q1 = 9.275 Q2 = 26.905 Q3 = 84.9425

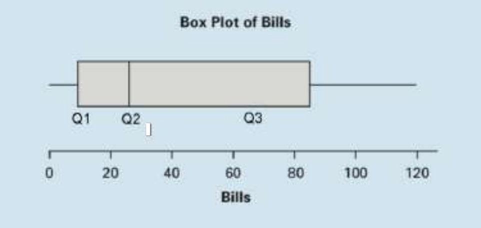
Interquartile range: Q3 - Q1 = 75.6675

To find the outliers:

- 1.5 x 75.6675 = 113.5013
- 9.275 - 113.5013 = 104.226
- 84.9425 + 113.5013 = 198.4438.

The Box Plot

The Box Plot shows us 5 statistics.



$Q1 = 9.275$ $Q2 = 26.905$ $Q3 = 84.9425$

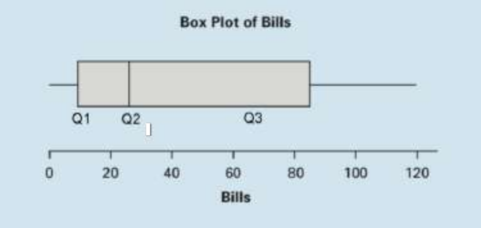
Interquartile range: Q3 - Q1 = 75.6675

To find the outliers:

- $1.5 \times 75.6675 = 113.5013$
- $9.275 - 113.5013 = \; 104.226$
- $84.9425 + 113.5013 = 198.4438.$

The Box Plot

The Box Plot shows us 5 statistics.



$$Q1 = 9.275 \; Q2 = 26.905 \; Q3 = 84.9425$$

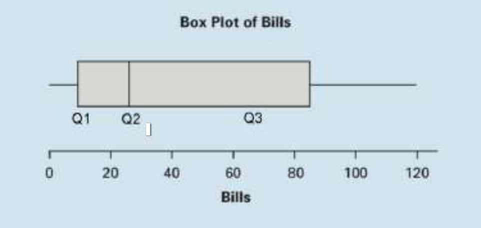
Interquartile range: Q3 - Q1 = 75.6675

To find the outliers:

- 1.5 x 75.6675 = 113.5013
- 9.275 - 113.5013 = 104.226
- 84.9425 + 113.5013 = 198.4438.

The Box Plot

The Box Plot shows us 5 statistics.



Q1 = 9.275 Q2 = 26.905 Q3 = 84.9425

Interquartile range: Q3 - Q1 = 75.6675

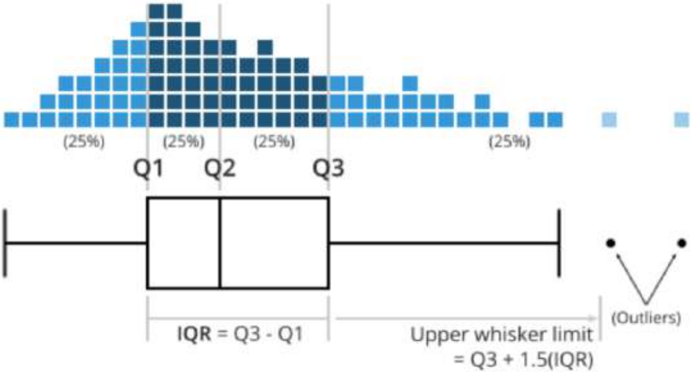
To find the outliers:

- 1.5 x 75.6675 = 113.5013
- 9.275 - 113.5013 = 104.226
- 84.9425 + 113.5013 = 198.4438.

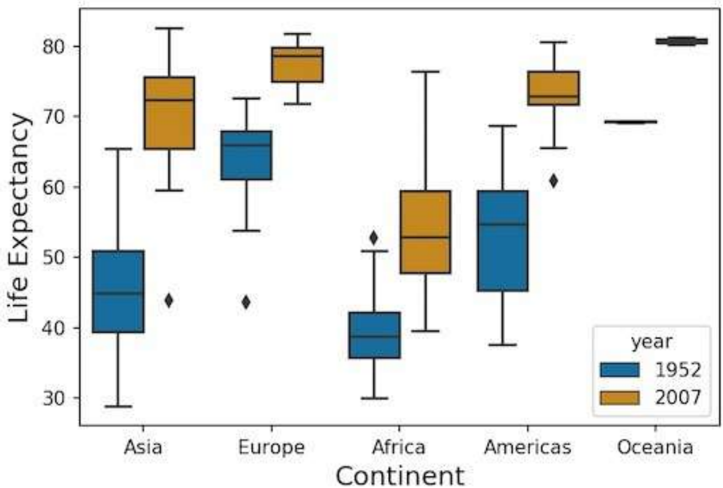
The whiskers extend outward to the smaller of 1.5 times the interquartile range or to the most extreme point that is not an outlier.

The Box Plot

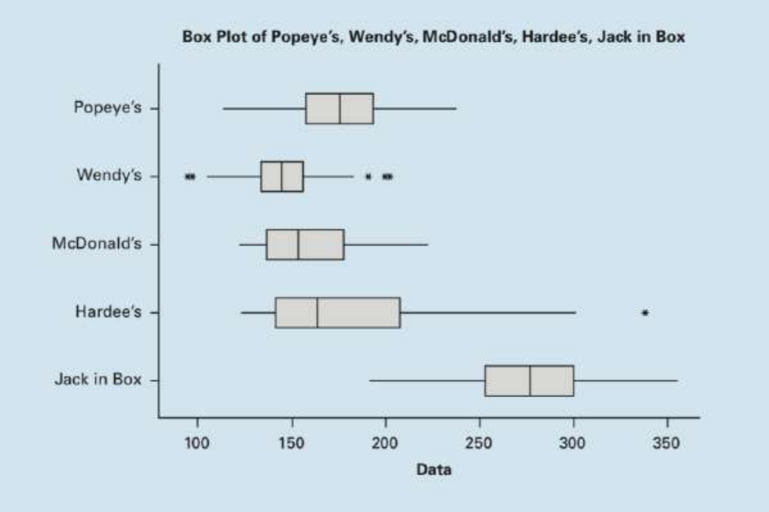
Box Plot and Histogram



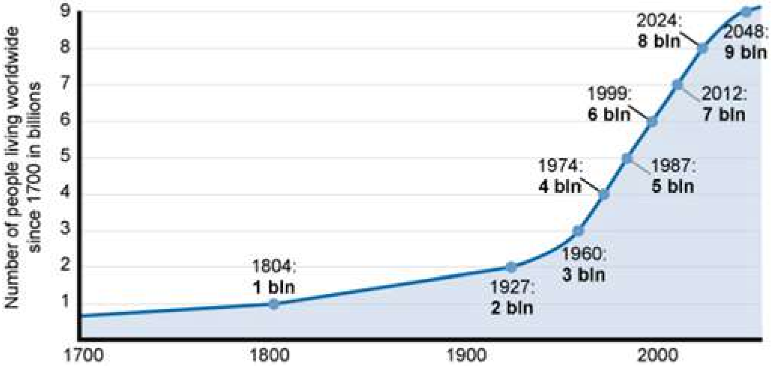
Life Expectancy by Continent



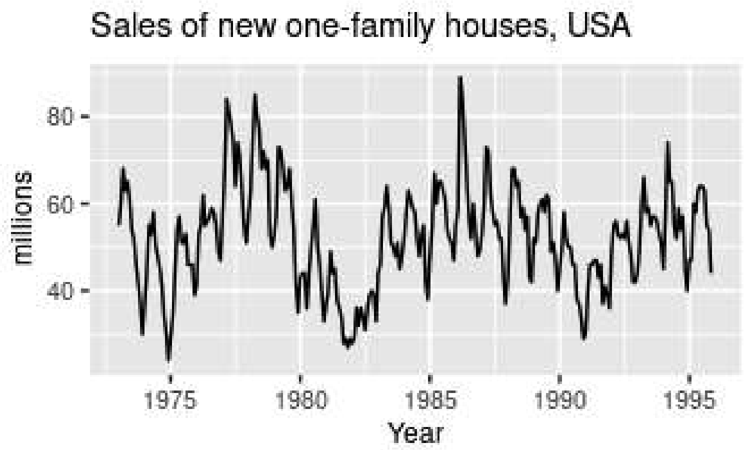
Delivery Times Fast Food Chains



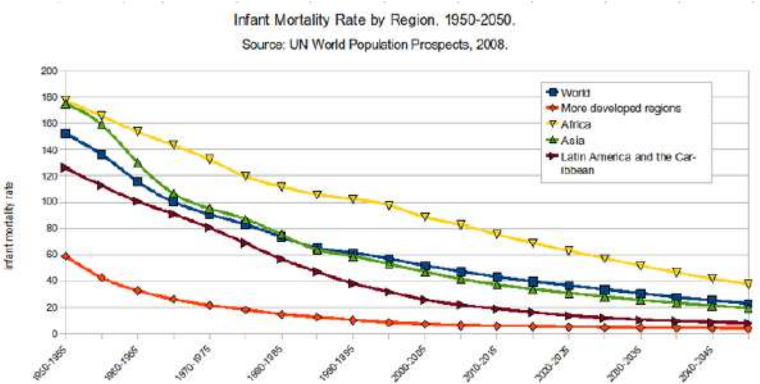
Positively Trending Line Chart



Line Chart without Trend



Negatively Trending Line Chart



1. Graphical Techniques for Interval Data Continued...
   - The Ogive Curve
   - Box Plots
   - Line Charts

2. Description of the Relationship Between Two Variables
   - Scatter Plot
   - Covariance
   - Correlation Coefficient

3. Practical Assignments in Python

Relationship between two variables

We often need to understand the relationship between two variables.

- Pricing departments want to know how the changes they make to prices affects their sales.
- Policy makers want to know how education affects crime rate in a country.
- Doctors want to know how daily exercise affects people's health.

To understand the relationship between two variables practitioners use Scatter Plots.

Relationship between two variables

We often need to understand the relationship between two variables.

- Pricing departments want to know how the changes they make to prices affects their sales.
- Policy makers want to know how education affects crime rate in a country.
- Doctors want to know how daily exercise affects people's health.

To understand the relationship between two variables practitioners use Scatter Plots.

Relationship between two variables

We often need to understand the relationship between two variables.

- Pricing departments want to know how the changes they make to prices affects their sales.
- Policy makers want to know how education affects crime rate in a country.
- Doctors want to know how daily exercise affects people's health.

To understand the relationship between two variables practitioners use Scatter Plots.

Relationship between two variables

We often need to understand the relationship between two variables.

- Pricing departments want to know how the changes they make to prices affects their sales.
- Policy makers want to know how education affects crime rate in a country.
- Doctors want to know how daily exercise affects people's health.

To understand the relationship between two variables practitioners use Scatter Plots.

Relationship between two variables

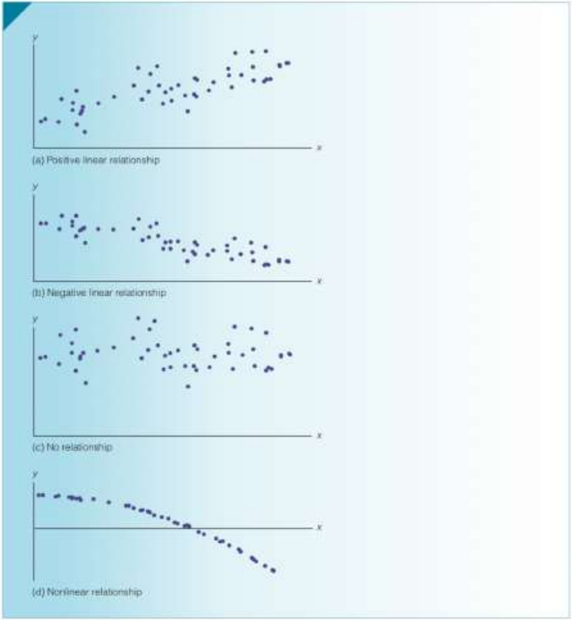
We often need to understand the relationship between two variables.

- Pricing departments want to know how the changes they make to prices affects their sales.
- Policy makers want to know how education affects crime rate in a country.
- Doctors want to know how daily exercise affects people's health.

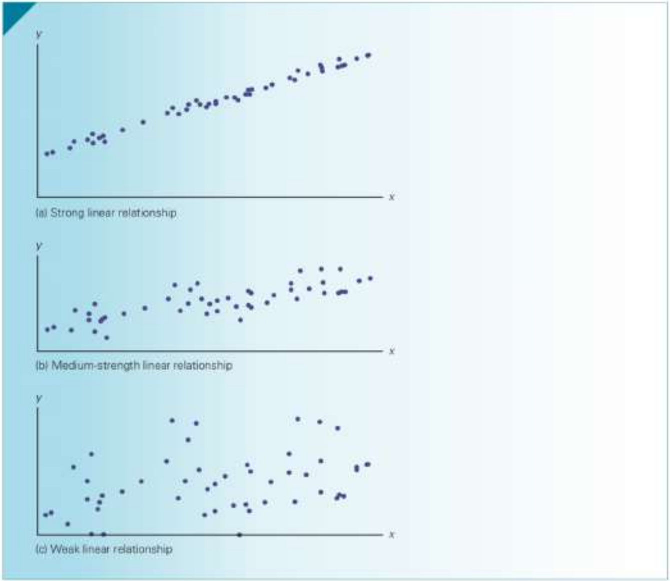
To understand the relationship between two variables practitioners use Scatter Plots.

Price of House versus Size

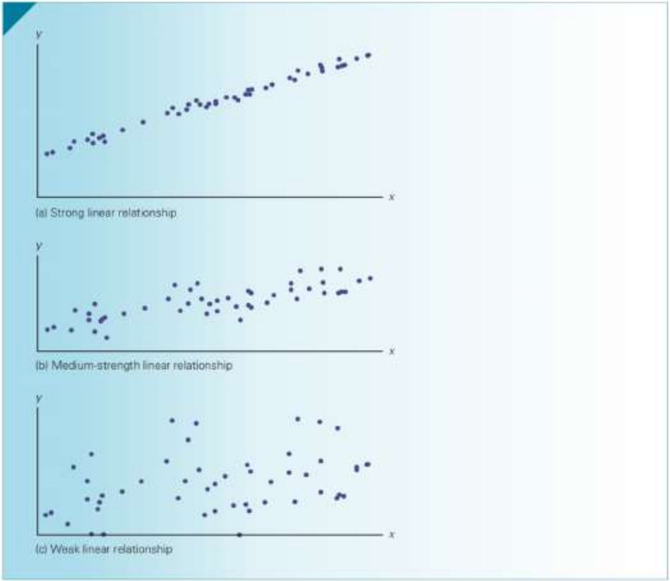
Relationship Direction & Type



Relationship Strength

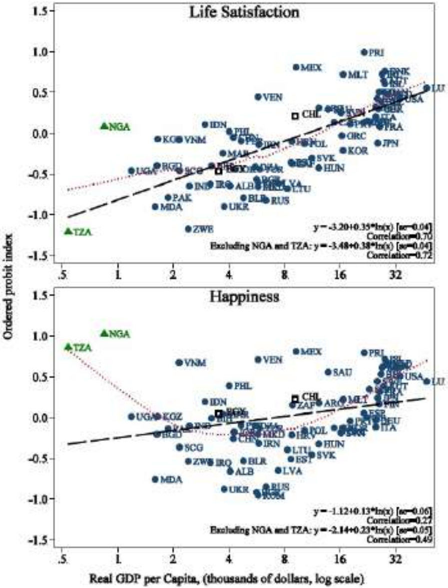


Relationship Strength

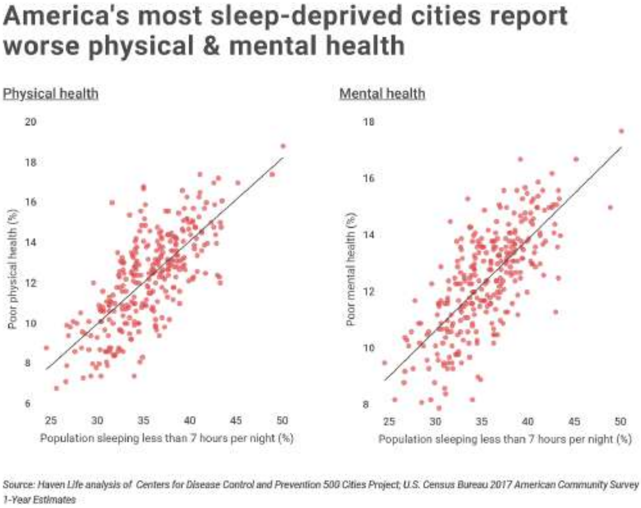


Correlation Examples

Figure 5. Subjective Well-Being and Real GDP per Capita: 1999-2004 World Values Survey



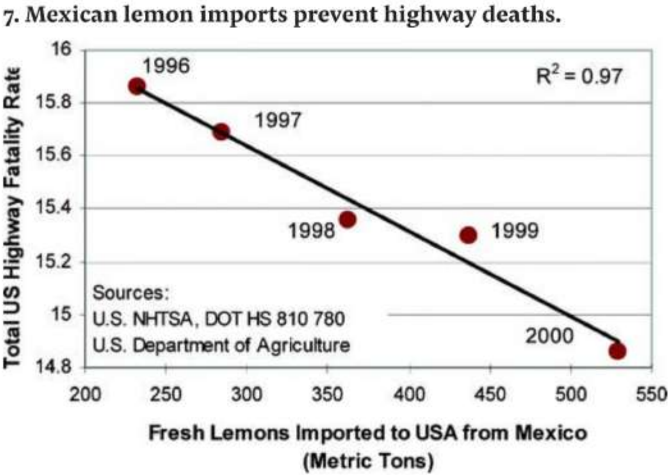
Correlation Examples



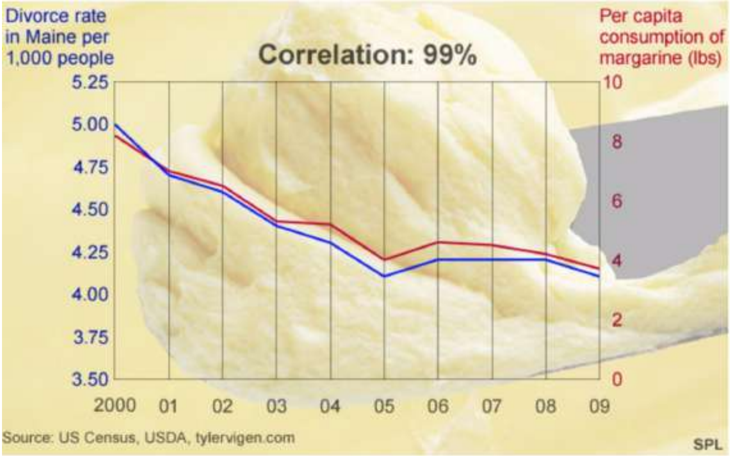
Strong correlation does not mean causality!

Price of shares Facebook versus Apple

Spurious Correlation



Spurious Correlation



Covariance

Covariance is a measure of the relationship between two random variables. The metric evaluates how much - to what extent - the variables change together. In other words, it is essentially a measure of the variance between two variables.

Covariance

Population covariance: $\sigma_{xy} = \dfrac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$

Sample covariance: $s_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$

Covariance Numerical Example

Set 1

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
2	13	–3	–7	21
6	20	1	0	0
7	27	2	7	14
$\bar{x} = 5$	$\bar{y} = 20$			$s_{xy} = 35/2 = 17.5$

Set 2

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
2	27	–3	7	–21
6	20	1	0	0
7	13	2	–7	–14
$\bar{x} = 5$	$\bar{y} = 20$			$s_{xy} = -35/2 = -17.5$

Set 3

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
2	20	–3	0	0
6	27	1	7	7
7	13	2	–7	–14
$\bar{x} = 5$	$\bar{y} = 20$			$s_{xy} = -7/2 = -3.5$

Covariance Shortcomings

We would like to know about the relationship between two variables:

- The sign of the relation
- The magnitude, as it shows the strength of the association

The magnitude is impossible to judge without additional statistics.

Solution: Correlation Coefficient

Covariance Shortcomings

We would like to know about the relationship between two variables:

- The sign of the relation
- The magnitude, as it shows the strength of the association

The magnitude is impossible to judge without additional statistics.

Solution: Correlation Coefficient

Covariance Shortcomings

We would like to know about the relationship between two variables:

- The sign of the relation
- The magnitude, as it shows the strength of the association

The magnitude is impossible to judge without additional statistics.

Solution: Correlation Coefficient

Covariance Shortcomings

We would like to know about the relationship between two variables:

- The sign of the relation
- The magnitude, as it shows the strength of the association

The magnitude is impossible to judge without additional statistics.

Solution: Correlation Coefficient

Covariance Shortcomings

We would like to know about the relationship between two variables:

- The sign of the relation
- The magnitude, as it shows the strength of the association

The magnitude is impossible to judge without additional statistics.

Solution: Correlation Coefficient

Correlation Coefficient

A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. The sign tells us about the direction of the relationship and the magnitude about its strength.

Coefficient of Correlation

Population coefficient of correlation: $\rho = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

Sample coefficient of correlation: $r = \dfrac{s_{xy}}{s_x s_y}$

Covariance Numerical Example

Set 1

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
2	13	–3	–7	21
6	20	1	0	0
7	27	2	7	14
$\bar{x} = 5$	$\bar{y} = 20$			$s_{xy} = 35/2 = 17.5$

Set 2

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
2	27	–3	7	–21
6	20	1	0	0
7	13	2	–7	–14
$\bar{x} = 5$	$\bar{y} = 20$			$s_{xy} = -35/2 = -17.5$

Set 3

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
2	20	–3	0	0
6	27	1	7	7
7	13	2	–7	–14
$\bar{x} = 5$	$\bar{y} = 20$			$s_{xy} = -7/2 = -3.5$

Correlation Coefficient Computation

$$\bar{x} = \frac{2+6+7}{3} = 5.0$$

$$\bar{y} = \frac{13+20+27}{3} = 20.0$$

$$s_x^2 = \frac{(2-5)^2+(6-5)^2+(7-5)^2}{3-1} = \frac{9+1+4}{2} = 7.0$$

$$s_y^2 = \frac{(13-20)^2+(20-20)^2+(27-20)^2}{3-1} = \frac{49+0+49}{2} = 49.0$$

The standard deviations are

$$s_x = \sqrt{7.0} = 2.65$$
$$s_y = \sqrt{49.0} = 7.00$$

The coefficients of correlation are:

Set 1: $r = \frac{s_{xy}}{s_x s_y} = \frac{17.5}{(2.65)(7.0)} = .943$

Set 2: $r = \frac{s_{xy}}{s_x s_y} = \frac{-17.5}{(2.65)(7.0)} = -.943$

Set 3: $r = \frac{s_{xy}}{s_x s_y} = \frac{-3.5}{(2.65)(7.0)} = -.189$

Correlation Coefficient Interpretation

$$-1 \leq r \leq +1 \quad \text{and} \quad -1 \leq \rho \leq +1$$

- -1 and 1 - perfectly linear relationship. A change in one variable is accompanied by a perfectly consistent change in the other.
- 0 - no linear relationship. As one variable increases, there is no tendency in the other variable to either increase or decrease.
- Between 0 and +1/-1 - there is a relationship, but the points dont all fall on a line.

Correlation Coefficient Interpretation

$$-1 \leq r \leq +1 \quad \text{and} \quad -1 \leq \rho \leq +1$$

- -1 and 1 - perfectly linear relationship. A change in one variable is accompanied by a perfectly consistent change in the other.
- 0 - no linear relationship. As one variable increases, there is no tendency in the other variable to either increase or decrease.
- Between 0 and +1/-1 - there is a relationship, but the points dont all fall on a line.

Correlation Coefficient Interpretation

$$-1 \leq r \leq +1 \qquad \text{and} \qquad -1 \leq \rho \leq +1$$

- -1 and 1 - perfectly linear relationship. A change in one variable is accompanied by a perfectly consistent change in the other.
- 0 - no linear relationship. As one variable increases, there is no tendency in the other variable to either increase or decrease.
- Between 0 and +1/-1 - there is a relationship, but the points dont all fall on a line.

Correlation Coefficient Interpretation

$$-1 \leq r \leq +1 \qquad \text{and} \qquad -1 \leq \rho \leq +1$$

- -1 and 1 - perfectly linear relationship. A change in one variable is accompanied by a perfectly consistent change in the other.
- 0 - no linear relationship. As one variable increases, there is no tendency in the other variable to either increase or decrease.
- Between 0 and $+1$/-1 - there is a relationship, but the points dont all fall on a line.

Let's get Started!

Access Google Colaboratory through your Gmail account